M.R.P. 1023
S.C. 11/225
31st December, 1956

71184

The serial correlations of daily mean temperatures at Kew Observatory

by

J. M. Craddock

(Meteorological Office)

AIR MINISTRY

METEOROLOGICAL RESEARCH COMMITTEE

At the time of issue this report has not been published. Any author desiring to quote this paper should refer to it as "A paper of the Meteorological Research Committee (London), M.R.P. No. 1023 by Mr. J. M. Craddock
A copy is available in the Library of the Meteorological Office."

M.R.P.1023
S.C.II/225
31st December, 1956

AIR MINISTRY

METEOROLOGICAL RESEARCH COMMITTEE

The serial correlations of daily mean temperatures at Kew Observatory

by
J. M. Craddock (Meteorological Office)

Summary

A study is made of the correlation between the daily mean temperature at Kew on each calendar day of the year with the temperature on each of the seven following days. It appears that the level of the correlations varies with the time of year. For any particular season the correlation decreases as the separation between the dates is increased, but generally remains positive for separations of up to seven days. The correlations are used to suggest a division of the year into natural seasons. Regression equations are given to enable the temperature on any day to be predicted in terms of those of the preceding days.

1. Introduction

It is well known that the mean daily air temperature near the earth's surface shows coherent departure from normal over periods of days or weeks. The present paper sets out to measure this coherence numerically. The direct and obvious method of doing so is to find the serial correlations of a series of daily mean temperature anomalies covering at least 50 to 100 consecutive days. This is open to the objection that data will extend over a considerable part of a year, while the correlations we are trying to measure may themselves have an annual variation.

The method actually used is to take a very long series of daily temperature anomalies (18,615 terms, covering 51 years) and to perform a separate analysis for each calendar day of the year. For example, for January 1, we find the mean temperature and standard deviation for the 51 years 1900 to 1950 and the serial correlations r_1 up to r_7 between the temperature on January 1 and that on each succeeding day up to January 8. A similar analysis is repeated every day of the year.

A computation of this size could not be carried out in a reasonable time without high speed computing facilities and the main part was done by punched card methods, using the Hollerith installation at Rothamsted Experimental Station by the courtesy of Dr. F. Yates.

2. Data and computation

Kew Meteorological Observatory is a station representative of much of South-East England at which meteorological observations of high quality have been maintained over many years. The data used covering the period 1900-1950 were taken from manuscript tabulations made at the observatory, giving separately for each day of each year the daily maximum and minimum temperature. These values were punched direct from the source documents on to Hollerith cards. The daily mean temperatures, the sums required to find mean values and the sums of squares and products required to give variances and covariances were all found mechanically using the Hollerith Tabulator without any further manual transcription. The final calculations to give means, standard deviations and correlation coefficients were carried out by hand and checked by repetition. A detailed description of the calculations and the steps taken to minimize the risk of errors is given in Appendix A.

/It

It should be mentioned that the daily mean temperature was estimated as $\frac{1}{2}$(max. + min.). While this method of estimation has its disadvantages as pointed out, for example, by Hoblyn (1928) it is still probably the best for general use, taking into account questions such as the availability of printed tabulations, and the time which would otherwise be spent on averaging data taken from original records. It is most unlikely that any of the main results of this paper would have been different had some other estimate of the daily mean temperature been used.

3. Results

The standard deviation of the daily mean temperature for each day of the year, and the correlation between the mean temperature on that day and the temperature on each of the seven following days, is shown diagrammatically in Fig.1. The results which number 2920 are rather numerous to present in tabular form, but can be made available to any investigator who wishes to use them. The salient features may be summarised as follows:-

3.1. The standard deviation of daily mean temperature

Considering first the lowest graph in Fig.1 it appears that besides a rather irregular fluctuation in value between one day and the next, this quantity shows a slow variation throughout the year. During the winter period from about November 20 to February 20 its average value is about 3.3°C, after which it falls to a value during March and April of about 2.7°C. From about May 8 to June 5 the value rises to about 3.0°C after which it falls to about 2.2°C during early August, and then rises steadily till it reaches its winter value in late November.

There is little doubt that these broad features are genuine, and not due to any accident of the statistics, but we cannot by elementary arguments assess the significance of the finer detail, such as the minor peak in mid-April, or the trough centred on December 24. To give the broad outlines concisely, the monthly mean values are given in Table 1.

It is hoped to carry out a more elaborate statistical analysis in another paper.

3.2. The variation of the serial correlations from month to month

If the remaining graphs in Figure 1 are considered in ascending order, they seem to differ more and more from what would be expected if the values of the serial correlations were independent of the time of year. We may put this impression into numerical form by using these data to estimate monthly values of the various coefficients.

The statistical distribution of a correlation coefficient departs a good deal from the normal, so that it is undesirable to average them as they stand. Instead each value in Figure 1 was converted into the z-statistic which is almost normally distributed and with 51 pairs of observations has variance 1/48. The monthly mean values of z are given in Table 2 and the corresponding values of r in Table 3.

We have to decide whether any of the monthly values of z in Table 2 depart significantly from the appropriate annual average. If consecutive values of z (or r) were independent, then the monthly mean value of z would have variance $1/48m$ where m is the number of days in the month. However, Figure 1 appears to confirm what mathematical reasoning would suggest, namely, that consecutive values of r are not independent. This fact reduces the number of degrees of freedom of the estimate of each monthly mean value of z, and we have to decide by how much. The adjustment has been discussed by, inter alia, Brooks (1948) and by Lewis and McIntosh (1952) and the latter authors give a formula for correcting m in terms of the serial correlations of the data in question. The calculation of several serial correlation coefficients for each set of 365 values of z would, however, have considerably increased the work put into this already laborious research so a simpler and cruder method of estimation was used, which is probably quite good enough for the purpose.

/When

When consecutive terms of a series are positively correlated, the expectation that a given term is followed by another which departs from normal in the same direction will be greater than it would be in a series of independent terms. The expectation of finding a long run of consecutive departures in the same sense will be greater, and hence the average length of run. Conversely, by finding the average length of runs of terms departing from normal in the same direction, we should be able to estimate the first serial correlation. The question has been discussed by Ogawara and others (1955) and his diagram for the estimation of r_1 is given with full working directions in Appendix B.

This method is very quick in use since the only operation required on the data is that of counting the runs of consecutive above or below normal values. The first serial correlation was estimated in this way for each of the sets of 365 values of z_1 up to z_7, the values being given in the last line but one of Table 2. On the further assumption that for these series the consecutive serial correlations are in geometrical progression we arrive at the values of the reducing factor E (given in Table 2) by which the actual number of observations in each sample are to be multiplied to find the number of effectively independent observations.

Using these values of E, and allowing for the months having 28, 30 or 31 days, the coefficients underlined in Table 3 appear to differ significantly at the 5 per cent level from the annual mean values.

All the February values are significantly higher than the annual mean, while all but one of those for November are significantly lower; and while the values of r_1 and r_2 for May differ little from the annual mean values, the May values of r_4 to r_7 are significantly below the annual values. While the statistical procedure used is only approximate, the consistency of the results makes it extremely probable that the observed variations between months are real and not due to any accident of the statistics. To give a numerical measure to this probability would mean much extra work which at this stage hardly seems justified.

In every month the serial correlations decrease as the interval between the days correlated is increased, but the monthly averages are still all positive for dates seven days apart.

4. The division of the year into "natural seasons"

Further examination of Figure 1 prompts the enquiry whether the calendar months are really the most natural way of subdividing the year. On a date for which values of r_1 to r_7 are comparatively high the temperature departure is more likely to persist than it is on dates with lower values of r_1 to r_7. A trough in the values of r_1 indicates a date or period when temperature anomalies are relatively unstable. We may distinguish seasons with undisturbed temperature regimes from others with more disturbed regimes and a break between two periods with high values of r is an obvious choice for the end of a "natural season". Lamb (1950) has made a similar subdivision but based on the persistence of weather types rather than that of temperature anomalies. Looking at the upper curves in Figure 1, for which the variations are more marked, and starting with the summer, the first real break covers the period September 14 to September 22. There is then a short relatively undisturbed spell until October 9 and then a disturbed temperature regime until December 3. Then, apart from short breaks about December 10, December 22 and January 9 the type remains relatively undisturbed till about March 12. The curve then shows an oscillation with seven peaks and six troughs in the 95 days until June 15 after which the temperature regime becomes progressively less disturbed, without major breaks until early September.

This subdivision may be compared with that devised by Lamb (1950) who divided the year into five seasons, the terminal dates being September 9, November 19, January 19, March 29 and June 17. Two of these dates, September 9 and June 17, agree fairly well with the division suggested above while a third, March 29, coincides with a well marked trough in the correlation graphs. The other dividing dates, November 19 and January 19, do not agree with any discernible features of the correlation graphs. From the present data it

would be more natural to consider autumn as running from September 15 to December 3, and winter as continuing from December 4 to March 12. Spring will then run from March 13 to June 15, and summer from June 16 to September 14.

5. Statistical forecasts

Given the correlations between the temperature anomaly on a particular day and those on several preceding days, a multiple regression equation can be found to give the best statistical forecast of the temperature in terms of the earlier days. This is an application of the method of Wiener (1950) but can readily be justified from first principles.

Suppose T'_n is the temperature anomaly on day n, and $T''_n(K)$ is the best possible forecast of T'_n obtainable as a linear combination of the anomalies for the K preceding days.

Let $$T''_n(K) = a_1 T'_{n-1} + a_2 T'_{n-2} \quad \cdots\cdots + a_K T'_{n-K} \qquad (1)$$

and $$T'_n - T''_n(K) < \epsilon_n(K) \qquad (2)$$

Then the problem is to choose $a_1 \cdots a_K$ to make the general run of values of $\epsilon_n(K)$ as small as possible: to do this we impose the condition that $\overline{\epsilon_n^2(K)}$ should be a minimum.

If the values of $a_1 \cdots a_K$ have been chosen to make $\epsilon_n(K)$ minimum, then $\epsilon_n(K)$ must be uncorrelated with $T'_{n-1}, T'_{n-2} \cdots T'_{n-K}$ because if it were correlated, positively or negatively, another set of values of $a_1 \cdots a_K$ could be found, allowing for the correlation which would have a lower mean value for $\epsilon_n^2(K)$.

Thus $$\overline{\epsilon_n(K)\,T'_{n-1}}\ ;\ \overline{\epsilon_n(K)\,T'_{n-2}} \cdots\cdots \overline{\epsilon_n(K)\,T'_{n-K}}$$

all tend to zero if summed over many cases.

But $$\epsilon_n(K) = T'_n - a_1 T'_{n-1} - a_2 T'_{n-2} \cdots\cdots - a_n T'_{n-K} \qquad (3)$$

so that multiplying in turn by $T'_{n-1}, T'_{n-2} \cdots T'_{n-K}$, summing over many cases, and dividing by $\sigma_{T'}^2$

$$\begin{aligned}
r_1 &= a_1 + a_2 r_1 + a_3 r_2 \quad\cdots\cdots + a_K r_{K-1} \\
r_2 &= a_1 r_1 + a_2 + a_3 r_1 \quad\cdots\cdots + a_K r_{K-2} \\
r_3 &= a_1 r_2 + a_2 r_1 + a_3 + a_4 r_1 \quad\cdots\cdots + a_K r_{K-3} \\
r_K &= a_1 r_{K-1} + a_2 r_{K-2} \quad\cdots\cdots + a_K
\end{aligned} \qquad (4)$$

Thus, given the values of $r_1 \cdots r_K$ we can find the corresponding values of $a_1 \cdots a_K$ by means of equations (4)

Further $$\epsilon_n^2 = \left(T'_n - a_1 T'_{n-1} \cdots\cdots - a_K T'_{n-K}\right)^2$$

and, when mean values are taken and equations (4) used, this becomes

$$\overline{\epsilon_n^2} = \overline{T'_n\left(T'_n - a_1 T'_{n-1} \cdots\cdots - a_n T'_{n-K}\right)} \qquad (5)$$

so that the relative residual error E_K is given by

$$E_K = \frac{\overline{\epsilon_n^2(K)}}{\sigma_{T'}^2} = \left(1 - a_1 r_1 - a_2 r_2 \cdots - a_K r_K\right) \qquad (6)$$

This enables us to estimate the proportional error remaining after the regression equation has been used and hence the value of the statistical forecast.

For large values of K the equations (4) can be solved by relaxation, or by the method described by Levinson (see Wiener (1950), Appendix B).

/For

For K = 2, 3 or 4 the algebraic solutions are simple enough to be useful, and are given in Table 4.

The choice of data on which to apply these equations deserves some consideration. If we work on correlation coefficients calculated for one particular day, we have the difficulties that the coefficients are subject to substantial sampling errors and that the relation between these errors and the sampling errors of the resulting regression coefficients is far from obvious. If we average values of z over a period, then the corresponding values of r will be statistically more stable, but will refer to the average conditions during the period which, unless we are careful in choosing the period, may differ considerably from the conditions on a particular day. We might try to set up some theoretical criterion to help us decide what to do, but it is probably more informative to see the results of the analysis of values of r relating to different trial periods.

First we consider the analysis of the values of r obtained from averaging z throughout the year, which statistically are the most stable obtainable from these data. The results are given in Table 5.

It appears that if the simplest formula is used (K = 1) the unpredicted variance amounts to 41½ per cent of the original value, while by using the 3-term regression equation (K = 3) the percentage residual variance is reduced to 41. The use of the 7-term formula, entailing much more work in calculation and in use produces a further improvement of 0.1 per cent. To put the matter differently, if we seek to forecast tomorrow's temperature by formulae applicable at all times of year, the mean square errors will have the values given in Table 6.

This shows that if we want a statistical formula to work all the year round, then the simple formula $T_n = 0.641\, T'_{n-1}$ is as good as anything more complicated. Incidentally, the standard of accuracy, as measured by the mean square error, is not very impressive. A synoptic forecaster using conventional methods might expect to do better.

Similar regression formulae may be based on the values of the serial correlations appropriate to each month of the year, and are given in Tables 7 to 18.

The regression equations take account of from one up to four consecutive terms, and it seems that, compared with the formulae based on annual values, the earlier terms may contribute a little more compared with the latest. Nevertheless, the improvement is not enough to suggest that the inclusion of still more terms is likely to justify the additional work of calculation. The use in forecasting routine of a statistical forecast obtained in this way is not likely to commend itself to synoptic meteorologists who would expect to do better by the exercise of personal skill. Nevertheless the statistical forecast may have value as a yardstick for research purposes, since it is objective and can be calculated quickly from past data, whereas the conditions under which a synoptic forecast is made cannot be reproduced properly after the event. For use as a yardstick the one-term regression formulae given in Tables 7 to 18 are probably as good as any others. Whatever formulae are used, the degree of success as measured by the E-value varies appreciably with the time of year.

6. It may of course be argued that rather than averaging the correlations, regression formulae should be calculated for each individual day and the significance of the regression coefficients judged from their stability from one day to the next. Whatever the theoretical advantages of this approach, it would involve more computation than could be undertaken at present, and the large sampling errors of the correlation coefficients might present a serious obstacle to the interpretation of the results. However, in Table 19 are given the correlations appropriate to non-overlapping 5-day periods covering the year. These should show the result of fairly short term variations in the coherence level while not being unduly affected by sampling errors.

/Examples

Examples of the analysis of these data are given in Tables 20 and 21. The period was chosen because r_1 was low and the simple one-term prediction formula was expected to be below the general level.

It appears that the improvement in E-value obtained by using a four-term regression formula is much larger than in the cases examined before. It is hard to say how much of this apparent improvement is due to the 5-day mean coefficients having larger sampling errors than the monthly, and the examples are given mainly to help other investigators to form their own judgments whether the use of more complicated regression equations is likely to produce any worthwhile increase in accuracy of prediction. My own view is that none of the more complicated formulae show enough improvement on the simple one-term formulae to justify their adoption.

Acknowledgments

Acknowledgments are due to Dr. Yates of the Statistical Department of Rothamsted Experimental Station for permission to use the Hollerith installation there, to Mr. M. J. R. Healy and Miss S. Cohen for advice and assistance in using it, and to Mrs. D. Gaskin and Miss N. P. Roberts who carried out most of the hand calculations.

References

Brooks, C.E.P.	(1948)	Met. Mag., London, Vol. 77, 1948, p.283.
Hoblyn, T.N.	(1928)	Quart.J.R. Met.Soc., Vol.54, 1928, p.183.
Lamb, H.H.	(1950)	Quart.J.R.Met.Soc., Vol.76, 1950, p.393.
Lewis, R.P.W. and McIntosh, D.H,	(1952)	Met.Mag., London, Vol.81, 1952, p.242.
Ogawara, M. and others	(1955)	Pap. Met. Geoph.,Tokyo, Vol.VI, No.1, p.87.
Wiener, N.	(1950)	Extrapolation, Interpolation and Smoothing of Stationary Time Series, New York. (John Wiley & Sons).

/Tables

TABLES

Table 1. Mean monthly values of variance $(°C)^2$ and standard deviation (°C)

	Variance	S.D.
January	11.56	3.40
February	10.80	3.29
March	8.03	2.83
April	7.60	2.76
May	8.40	2.90
June	7.06	2.66
July	6.61	2.57
August	5.42	2.33
September	6.64	2.58
October	7.65	2.77
November	9.43	3.07
December	10.89	3.30
Annual	8.328	2.886

Table 2. Monthly averages of z

	z_1	z_2	z_3	z_4	z_5	z_6	z_7
January	1.0474	0.6287	0.4974	0.3835	0.3042	0.2455	0.2032
February	1.1604	0.7771	0.6154	0.5063	0.4386	0.3775	0.3254
March	1.0077	0.6410	0.4548	0.3384	0.2890	0.2455	0.2213
April	1.0183	0.6020	0.3957	0.2823	0.2037	0.1273	0.1067
May	1.0455	0.6052	0.3694	0.2287	0.1290	0.0516	0.0126
June	1.0033	0.5947	0.4003	0.2907	0.2113	0.1707	0.1483
July	0.9900	0.6123	0.4223	0.3194	0.2606	0.2348	0.2348
August	0.9955	0.6232	0.4690	0.3974	0.3474	0.3250	0.3127
September	0.9490	0.5563	0.4107	0.3343	0.2783	0.2470	0.2050
October	0.9329	0.5426	0.3903	0.2945	0.2281	0.1752	0.1310
November	0.9180	0.5030	0.3320	0.2183	0.1500	0.1053	0.0743
December	1.0119	0.5919	0.4261	0.3381	0.2852	0.2552	0.2400
Ann. mean	1.0058	0.6056	0.4310	0.3253	0.2595	0.2117	0.1832
First serial correlation of series of z values	0.227	0.535	0.573	0.564	0.641	0.705	0.627
Reducing factor E	0.6774	0.3341	0.3014	0.3089	0.2467	0.1994	0.2575

Table 3. Monthly mean values of r

	r_1	r_2	r_3	r_4	r_5	r_6	r_7
January	0.7806	0.5574	0.4598	0.3653	0.2950	0.2402	0.2003
February	0.8210	0.6510	0.5476	0.4668	0.4128	0.3601	0.3140
March	0.7649	0.5656	0.4260	0.3257	0.2812	0.2402	0.2175
April	0.7691	0.5385	0.3765	0.2748	0.2012	0.1263	0.1066
May	0.7799	0.5406	0.3531	0.2251	0.1283	0.0520	0.0130
June	0.7629	0.5335	0.3799	0.2831	0.2079	0.1694	0.1469
July	0.7574	0.5455	0.3986	0.3086	0.2552	0.2308	0.2308
August	0.7595	0.5532	0.4374	0.3774	0.3337	0.3140	0.3032
September	0.7393	0.5050	0.3893	0.3221	0.2711	0.2421	0.2022
October	0.7320	0.4953	0.3714	0.2867	0.2241	0.1732	0.1303
November	0.7249	0.4645	0.3203	0.2146	0.1489	0.1046	0.0730
December	0.7666	0.5313	0.4020	0.3257	0.2775	0.2496	0.2355
Ann. mean	0.7641	0.5413	0.4062	0.3140	0.2534	0.2089	0.1810

Table 4. Algebraic solutions for regression coefficients

$K = 2$	$K = 3$
$a_1 = r_1(1-a_2) = \frac{r_1(1-r_2)}{1 - r_1^2}$ $a_2 = \frac{r_2 - r_1^2}{1 - r_1^2}$	$a_1 + a_3 = \frac{r_1 + r_3 - 2r_1r_2}{1 + r_2 - 2r_1^2}$ $a_1 - a_3 = \frac{r_1 - r_3}{1 - r_2}$ $a_2 = \frac{r_2(1+r_2) - r_1(r_1 + r_3)}{1 + r_2 - 2r_1^2}$

$K = 4$

$$a_1 + a_4 = \frac{(r_2 + r_3)(r_1 + r_2) - (r_1 + r_4)(1 + r_1)}{(r_1 + r_2)^2 - (1 + r_1)(1 + r_3)}$$

$$a_1 - a_4 = \frac{(r_2 - r_3)(r_1 - r_2) - (r_1 - r_4)(1 - r_1)}{(r_1 - r_2)^2 - (1 - r_1)(1 - r_3)}$$

$$a_2 + a_3 = \frac{(r_1 + r_4)(r_1 + r_2) - (r_2 + r_3)(1 + r_3)}{(r_1 + r_2)^2 - (1 + r_1)(1 + r_3)}$$

$$a_2 - a_3 = \frac{(r_1 - r_4)(r_1 - r_2) - (r_2 - r_3)(1 - r_3)}{(r_1 - r_2)^2 - (1 - r_1)(1 - r_3)}$$

Table 5. Prediction formulae based on annual mean values of r_1, r_2 -- r_7

K	Formula	E
1	$T''_n = 0.7641\ T'_{n-1}$	0.4162
2	$T''_n = 0.8421\ T'_{n-1} - 0.1021\ T'_{n-2}$	0.4119
3	$T''_n = 0.8491\ T'_{n-1} - 0.1603\ T'_{n-2} + 0.0689\ T'_{n-3}$	0.4100
4	$T''_n = 0.8486\ T'_{n-1} - 0.1593\ T'_{n-2} + 0.0627\ T'_{n-3} + 0.0077\ T'_{n-4}$	0.4099
7	$T''_n = 0.8479\ T'_{n-1} - 0.1615\ T'_{n-2} + 0.0673\ T'_{n-3} - 0.0186\ T'_{n-4} + 0.0245\ T'_{n-5} - 0.0130\ T'_{n-6} + 0.0272\ T'_{n-7}$	0.4094

/Table 6.

Table 6. Mean square errors of various statistical forecasts applicable throughout the year

Basis of forecast	σ_E (°C)
Climatic normal: $T''_n = 0$	2.89
$K = 1$: $T''_n = 0.7641\ T'_{n-1}$	1.87
$K = 7$: $T''_n = 0.8479\ T'_{n-1} - 0.1615\ T'_{n-2} + 0.0673\ T'_{n-3} - 0.0186\ T'_{n-4} + 0.0245\ T'_{n-5} - 0.0130\ T'_{n-6} + 0.0272\ T'_{n-7}$	1.86

Table 7. Prediction formulae based on January mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7806\ T'_{n-1}$	0.3907
2	$T''_n = 0.8843\ T'_{n-1} - 0.1328\ T'_{n-2}$	0.3837
3	$T''_n = 0.9087\ T'_{n-1} - 0.2957\ T'_{n-2} + 0.1839\ T'_{n-3}$	0.3709
4	$T''_n = 0.9242\ T'_{n-1} - 0.3209\ T'_{n-2} + 0.2613\ T'_{n-3} - 0.0845\ T'_{n-4}$	0.3683

Table 8. Prediction formulae based on February mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.8210\ T'_{n-1}$	0.3260
2	$T''_n = 0.8788\ T'_{n-1} - 0.0706\ T'_{n-2}$	0.3245
3	$T''_n = 0.8861\ T'_{n-1} - 0.1611\ T'_{n-2} + 0.1027\ T'_{n-3}$	0.3212
4	$T''_n = 0.8853\ T'_{n-1} - 0.1597\ T'_{n-2} + 0.0975\ T'_{n-3} + 0.0063\ T'_{n-4}$	0.3209

Table 9. Prediction formulae based on March mean value of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7649\ T'_{n-1}$	0.4149
2	$T''_n = 0.8009\ T'_{n-1} - 0.0470\ T'_{n-2}$	0.4140
3	$T''_n = 0.8019\ T'_{n-1} - 0.0642\ T'_{n-2} + 0.0217\ T'_{n-3}$	0.4137
4	$T''_n = 0.8018\ T'_{n-1} - 0.0638\ T'_{n-2} + 0.0140\ T'_{n-3} + 0.0095\ T'_{n-4}$	0.4137

/Table 10.

Table 10. Prediction formulae based on April mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7691\ T'_{n-1}$	0.4085
2	$T''_n = 0.8688\ T'_{n-1} - 0.1297\ T'_{n-2}$	0.4016
3	$T''_n = 0.8716\ T'_{n-1} - 0.1480\ T'_{n-2} + 0.0209\ T'_{n-3}$	0.4015
4	$T''_n = 0.8706\ T'_{n-1} - 0.1443\ T'_{n-2} + 0.0011\ T'_{n-3} + 0.0259\ T'_{n-4}$	0.4014

Table 11. Prediction formulae based on May mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7799\ T'_{n-1}$	0.3918
2	$T''_n = 0.9144\ T'_{n-1} - 0.1725\ T'_{n-2}$	0.3802
3	$T''_n = 0.9115\ T'_{n-1} - 0.1567\ T'_{n-2} - 0.0175\ T'_{n-3}$	0.3800
4	$T''_n = 0.9115\ T'_{n-1} - 0.1561\ T'_{n-2} - 0.0211\ T'_{n-3} + 0.0043\ T'_{n-4}$	0.3800

Table 12. Prediction formulae based on June mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7629\ T'_{n-1}$	0.4180
2	$T''_n = 0.8514\ T'_{n-1} - 0.1160\ T'_{n-2}$	0.4124
3	$T''_n = 0.8555\ T'_{n-1} - 0.1453\ T'_{n-2} + 0.0345\ T'_{n-3}$	0.4117
4	$T''_n = 0.8551\ T'_{n-1} - 0.1425\ T'_{n-2} + 0.0153\ T'_{n-3} + 0.0225\ T'_{n-4}$	0.4114

Table 13. Prediction formulae based on July mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7574\ T'_{n-1}$	0.4263
2	$T''_n = 0.8075\ T'_{n-1} - 0.0662\ T'_{n-2}$	0.4245
3	$T''_n = 0.8087\ T'_{n-1} - 0.0816\ T'_{n-2} + 0.0193\ T'_{n-3}$	0.4243
4	$T''_n = 0.8081\ T'_{n-1} - 0.0787\ T'_{n-2} + 0.0114\ T'_{n-3} + 0.0379\ T'_{n-4}$	0.4236

/Table 14.

Table 14. Prediction formulae based on August mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7595\ T'_{n-1}$	0.4232
2	$T''_n = 0.8019\ T'_{n-1} - 0.0558\ T'_{n-2}$	0.4219
3	$T''_n = 0.8067\ T'_{n-1} - 0.1247\ T'_{n-2} + 0.0859\ T'_{n-3}$	0.4187
4	$T''_n = 0.8011\ T'_{n-1} - 0.1163\ T'_{n-2} + 0.0313\ T'_{n-3} + 0.0675\ T'_{n-4}$	0.4167

Table 15. Prediction formulae based on September mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7393\ T'_{n-1}$	0.4534
2	$T''_n = 0.8072\ T'_{n-1} - 0.0918\ T'_{n-2}$	0.4496
3	$T''_n = 0.8173\ T'_{n-1} - 0.1807\ T'_{n-2} + 0.1101\ T'_{n-3}$	0.4442
4	$T''_n = 0.8140\ T'_{n-1} - 0.1745\ T'_{n-2} + 0.0844\ T'_{n-3} + 0.0307\ T'_{n-4}$	0.4435

Table 16. Prediction formulae based on October mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7320\ T'_{n-1}$	0.4642
2	$T''_n = 0.7958\ T'_{n-1} - 0.0872\ T'_{n-2}$	0.4607
3	$T''_n = 0.8037\ T'_{n-1} - 0.1584\ T'_{n-2} + 0.0891\ T'_{n-3}$	0.4571
4	$T''_n = 0.8031\ T'_{n-1} - 0.1577\ T'_{n-2} + 0.0867\ T'_{n-3} + 0.0031\ T'_{n-4}$	0.4571

Table 17. Prediction formulae based on November mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7249\ T'_{n-1}$	0.4745
2	$T''_n = 0.8181\ T'_{n-1} - 0.1286\ T'_{n-2}$	0.4667
3	$T''_n = 0.8274\ T'_{n-1} - 0.1872\ T'_{n-2} + 0.0718\ T'_{n-3}$	0.4642
4	$T''_n = 0.8299\ T'_{n-1} - 0.1935\ T'_{n-2} + 0.0993\ T'_{n-3} - 0.0335\ T'_{n-4}$	0.4637

Table 18. Prediction formulae based on December mean values of r_1 to r_7

K	Formula	E
1	$T''_n = 0.7666\ T'_{n-1}$	0.4123
2	$T''_n = 0.8715\ T'_{n-1} - 0.1368\ T'_{n-2}$	0.4046
3	$T''_n = 0.8861\ T'_{n-1} - 0.2309\ T'_{n-2} + 0.1083\ T'_{n-3}$	0.3999
4	$T''_n = 0.8839\ T'_{n-1} - 0.2257\ T'_{n-2} + 0.0897\ T'_{n-3} + 0.0229\ T'_{n-4}$	0.3995

Table 20. Prediction formulae based on mean values of r_1 to r_7 for the 5 days January 5 - 9

K	Formula	E
1	$T''_n = 0.6973\ T'_{n-1}$	0.5138
2	$T''_n = 0.7692\ T'_{n-1} - 0.1030\ T'_{n-2}$	0.5082
3	$T''_n = 0.7973\ T'_{n-1} - 0.3133\ T'_{n-2} + 0.2732\ T'_{n-3}$	0.4704
4	$T''_n = 0.8010\ T'_{n-1} - 0.3174\ T'_{n-2} + 0.2837\ T'_{n-3} - 0.0130\ T'_{n-4}$	0.4701

Table 21. Prediction formulae based on mean values of r_1 to r_7 for the 5 days January 10 - 14

K	Formula	E
1	$T''_n = 0.7192\ T'_{n-1}$	0.4828
2	$T''_n = 0.9186\ T'_{n-1} - 0.2771\ T'_{n-2}$	0.4455
3	$T''_n = 0.9605\ T'_{n-1} - 0.4163\ T'_{n-2} + 0.1515\ T'_{n-3}$	0.4354
4	$T''_n = 0.9825\ T'_{n-1} - 0.4749\ T'_{n-2} + 0.2865\ T'_{n-3} - 0.1411\ T'_{n-4}$	0.4265

/Table 19.

Table 19. Correlations and corresponding z values for 5-day periods

Period	z_1	r_1	z_2	r_2	z_3	r_3	z_4	r_4	z_5	r_5	z_6	r_6	z_7	r_7
1	.862	.697	.464	.433	.424	.400	.386	.368	.280	.273	.262	.256	.370	.354
2	.906	.719	.404	.383	.224	.220	.100	.100	.066	.066	.136	.135	.176	.174
3	1.032	.775	.656	.576	.516	.475	.448	.420	.404	.383	.296	.288	.182	.180
4	1.124	.809	.750	.635	.552	.573	.534	.488	.374	.357	.266	.260	.152	.151
5	1.246	.847	.768	.646	.570	.515	.478	.445	.396	.377	.300	.291	.300	.291
6	1.088	.796	.694	.601	.612	.546	.402	.382	.284	.277	.214	.211	.148	.147
7	1.180	.827	.744	.632	.542	.495	.438	.412	.398	.378	.368	.352	.314	.304
8	1.272	.854	.960	.744	.754	.638	.544	.496	.484	.449	.436	.410	.380	.363
9	1.190	.831	.820	.675	.682	.593	.616	.548	.484	.449	.404	.383	.334	.322
10	1.182	.828	.712	.612	.500	.462	.394	.375	.334	.322	.278	.271	.216	.213
11	1.052	.783	.710	.611	.584	.526	.480	.446	.422	.399	.354	.340	.294	.286
12	1.100	.800	.822	.676	.728	.622	.624	.554	.548	.499	.442	.415	.396	.377
13	1.016	.768	.610	.544	.386	.368	.356	.342	.356	.342	.306	.297	.264	.258
14	.792	.660	.420	.397	.302	.293	.204	.201	.230	.226	.236	.232	.238	.234
15	.924	.728	.536	.490	.294	.286	.108	.108	.092	.092	.052	.052	.012	.012
16	1.052	.783	.672	.586	.516	.475	.428	.404	.282	.275	.192	.190	.164	.163
17	1.124	.809	.814	.672	.594	.533	.488	.453	.462	.432	.496	.459	.512	.472
18	1.034	.776	.678	.590	.454	.425	.292	.284	.238	.234	.216	.213	.216	.213
19	.912	.722	.482	.448	.358	.343	.312	.302	.244	.239	.156	.155	.142	.141
20	.952	.741	.514	.473	.298	.289	.180	.178	.086	.086	.122	.121	.172	.170
21	1.232	.843	.802	.665	.622	.553	.518	.476	.460	.430	.314	.304	.114	.114
22	1.020	.770	.580	.523	.392	.373	.306	.297	.234	.230	.166	.164	.208	.205
23	1.006	.764	.588	.528	.342	.329	.174	.172	.036	.036	.008	.008	-.040	-.040
24	.904	.718	.462	.432	.250	.245	.182	.180	.162	.161	.062	.062	.042	.042
25	.980	.753	.608	.543	.348	.335	.142	.141	-.014	-.014	-.080	-.080	-.022	-.022
26	.986	.756	.510	.470	.328	.317	.238	.234	.136	.135	.002	.002	-.098	-.098
27	1.218	.839	.804	.666	.532	.487	.360	.345	.278	.271	.144	143	.074	.074
28	1.042	.779	.528	.484	.276	.269	.140	.139	.020	.020	.022	.022	.040	.040
29	1.136	.813	.690	.598	.394	.375	.164	.163	.020	.020	-.096	-.096	-.184	-.182
30	1.000	.762	.630	.558	.420	.397	.356	.342	.312	.302	.280	.273	.290	.282

Table 19. (Continued)

Period	z_1	r_1	z_2	r_2	z_3	r_3	z_4	r_4	z_5	r_5	z_6	r_6	z_7	r_7
31	1.056	.784	.642	.566	.426	.402	.262	.256	.122	.121	.014	.014	.030	.030
32	1.082	.794	.636	.562	.464	.433	.344	.331	.252	.247	.218	.215	.122	.121
33	.988	.757	.546	.498	.390	.371	.296	.288	.226	.222	.224	.220	.208	.205
34	.950	.740	.588	.528	.334	.322	.218	.215	.106	.106	.058	.058	.070	.070
35	.920	.726	.540	.493	.372	.356	.264	.258	.226	.222	.150	.149	.142	.141
36	.990	.757	.610	.544	.460	.430	.360	.345	.320	.310	.276	.269	.256	.251
37	.962	.745	.558	.506	.362	.347	.276	.269	.252	.247	.236	.232	.236	.232
38	1.108	.803	.676	.589	.412	.390	.338	.326	.308	.299	.282	.275	.236	.232
39	.878	.705	.518	.476	.368	.352	.250	.245	.190	.188	.186	.184	.148	.147
40	.942	.736	.570	.515	.338	.326	.308	.299	.286	.278	.290	.282	.302	.293
41	1.104	.802	.766	.645	.586	.527	.404	.383	.264	.258	.208	.205	.234	.230
42	.962	.745	.574	.518	.456	.427	.360	.345	.284	.277	.234	.230	.234	.230
43	1.008	.765	.660	.578	.404	.383	.306	.297	.292	.284	.252	.247	.200	.197
44	1.016	.768	.704	.607	.566	.512	.504	.465	.376	.359	.370	.354	.320	.310
45	.988	.757	.594	.533	.460	.430	.386	.368	.394	.375	.328	.317	.298	.289
46	1.010	.766	.638	.564	.498	.461	.414	.392	.310	.300	.290	.282	.324	.313
47	.964	.746	.512	.472	.370	.354	.358	.343	.342	.329	.284	.277	.280	.273
48	1.010	.766	.664	.581	.500	.462	.374	.357	.322	.311	.356	.324	.338	.326
49	1.004	.763	.668	.584	.546	.498	.430	.405	.346	.333	.352	.338	.330	.319
50	.986	.756	.612	.546	.494	.457	.470	.438	.508	.468	.500	.462	.418	.395
51	.882	.707	.496	.459	.390	.371	.388	.370	.376	.359	.342	.329	.304	.295
52	.920	.726	.438	.412	.278	.271	.204	.201	.130	.129	.116	.115	.160	.159
53	.920	.726	.516	.475	.340	.327	.214	.211	.094	.094	.044	.044	-.034	-.034
54	.972	.750	.608	.543	.462	.439	.342	.329	.260	.254	.202	.199	.136	.135
55	1.002	.762	.666	.582	.442	.415	.298	.289	.244	.239	.246	.241	.246	.241
56	.960	.744	.596	.534	.584	.526	.544	.496	.460	.430	.368	.352	.262	.256
57	.868	.700	.424	.400	.256	.251	.228	.224	.258	.252	.268	.262	.232	.228
58	.944	.737	.484	.449	.308	.299	.232	.228	.172	.170	.072	.072	.038	.038
59	.844	.688	.444	.417	.290	.282	.242	.237	.254	.249	.248	.243	.190	.188
60	.976	.751	.618	.550	.468	.437	.278	.271	.090	.090	-.006	-.006	-.034	-.034

Table 19. (Continued)

Period	z_1	r_1	z_2	r_2	z_3	r_3	z_4	r_4	z_5	r_5	z_6	r_6	z_7	r_7
61	.914	.723	.514	.473	.298	.289	.190	.188	.144	.143	.100	.100	.128	.127
62	.826	.678	.436	.410	.264	.258	.112	.112	.026	.026	.016	.016	-.010	-.010
63	.924	.728	.574	.518	.358	.343	.270	.264	.186	.184	.148	.147	.174	.172
64	.946	.738	.450	.422	.298	.289	.202	.199	.152	.151	.106	.106	.068	.068
65	.926	.729	.508	.468	.332	.320	.218	.215	.140	.139	.090	.090	.008	.008
66	.972	.750	.564	.511	.432	.407	.336	.324	.266	.260	.162	.161	.100	.100
67	.958	.743	.552	.502	.338	.326	.186	.184	.086	.086	.110	.110	.126	.125
68	.918	.725	.478	.445	.354	.340	.268	.262	.190	.188	.096	.096	.016	.016
69	1.036	.776	.666	.582	.542	.495	.450	.422	.370	.354	.392	.373	.348	.335
70	1.166	.823	.730	.623	.522	.479	.374	.357	.298	.289	.224	.220	.104	.104
71	.904	.718	.456	.427	.366	.350	.312	.302	.256	.251	.184	.182	.168	.166
72	1.096	.799	.650	.572	.410	.388	.286	.278	.226	.222	.134	.133	.084	.084
73	1.054	.783	.630	.558	.428	.404	.360	.345	.434	.409	.302	.293	.486	.451

Appendix A

Data and computation

The original tabulations made at Kew Observatory gave separately for each day of each year from 1900 to 1950 the daily maximum and minimum temperature. These were given in degrees Absolute to the nearest tenth of a degree.

Each temperature (omitting the initial figure, nearly always 2) formed a 3-figure group. The maxima were punched on one pack of Hollerith cards and the minima in corresponding positions on another. In each pack, data for the first 16 days of each month are on one card and those for the rest of the month on another, so that each of the two packs covering the 51 years contains 1224 cards. The punching was verified by the normal Hollerith procedure and finally the entire contents of each card were listed, using the Hollerith Tabulator, and the printed version was checked back against the source documents. There is no reasonable doubt that the data punched on the cards which form the basis of all the calculations are accurately copied from the record made at the Observatory.

These two packs of Hollerith cards bear in corresponding positions the maximum and minimum temperatures for the same day. The packs were interleaved and the temperatures added on a Hollerith Tabulator, with a break of control after every second card, the totals being summary punched into corresponding positions on a blank pack of cards. This third pack effectively gives the daily mean temperature in units of $\frac{1}{2}$ degree Absolute. February 29 was ignored when it occurred.

Most of the work in finding variances and correlations consists in finding the sums of squares and products. If the daily mean temperatures are denoted by T_1, T_2, etc., then the sums ΣT_1^2 and products $\Sigma T_1 T_2$, $\Sigma T_1 T_3$ up to $\Sigma T_1 T_7$ were found on the Hollerith Tabulator by progressive digiting. Here T_1 is supposed to represent the temperature on any particular calendar date, and the summations to extend over the 51 years 1900-1950. In the progressive digiting each 'summation' such as $\Sigma T_1 T_2$ is obtained in two different ways and thus checked for accuracy, but the sums of squares are not so checked. However, on account of the way the calculation was arranged each result had a large non-significant constant added to it, such that any mistakes in the non-significant digits would probably be obvious. Apart from known machine failures no mistakes seem to have occurred in the non-significant digits so that the number of mistakes among the significant figures is probably very small.

From each printed result the constant (3,511,485) was subtracted, and the variances, covariances and serial correlations were found by the formulae

$$V(T_1) = \frac{n\Sigma T_1^2 - (\Sigma T_1)^2}{n(n-1)}$$

$$\mathrm{cov}(T_1 T_2) = \frac{n\Sigma T_1 T_2 - (\Sigma T_1)(\Sigma T_2)}{n(n-1)}$$

$$r_k(T_1) = \frac{\mathrm{cov}\, T_1 T_{1+k}}{\sqrt{V(T_1)\, V(T_{1+k})}}$$

These steps were done on a hand or electric calculating machine, the results being checked by repetition. In three cases out of 2455, the final calculation gave a value of r exceeding unity. It was confirmed that no error had occurred in the hand part of the calculation, so that the inconsistency must be present in the printed figures output from the Tabulator. At first sight the fault is more likely to arise in the sums of squares which are unchecked, rather than in the sums of products which are checked, but it seems very unlikely that an error in a sum of squares should produce an impossible

/value

value for only one of the fourteen correlations into which it enters. The checking of the sums of products is by comparing two columns of printed figures, and perhaps the most likely explanation is that in these three cases a discrepancy escaped the notice of the scrutineer. These errors were only discovered months after the completion of the Hollerith part of the calculation, and it would have been quite awkward to have recomputed the values, so the mean values for the appropriate month have been used in Fig. 1 instead.

Appendix B

A diagram for simple estimation of a correlation coefficient

If two variables are positively correlated we would expect that the number of cases in which the signs of corresponding deviations agree would exceed the number in which the signs disagree. The relation between the correlation coefficient *r* and the proportion of cases with like signs *p* is discussed by Ogawara and others (1955) on the assumption that both the variables are normally distributed. Fig. 2 is a simplified version of the diagram given by Ogawara, showing the unique relation between *p* and *r* when the sample considered is infinitely large, and the limits within which *r* is expected to lie (at the 10 per cent confidence level) if *p* is found from a sample of 500, 100 or 20 pairs.

In dealing with one autocorrelated series let *p* be the probability that a term which departs from normal in one direction is followed by another which departs in the same direction. *p* is assumed not to depend on whether the departure is positive or negative or on the sequence of previous anomalies. Then the probability that a chosen term is followed by exactly *n* more with anomalies of the same sign is $p^n(1-p)$, and the probability that it is preceded by exactly *m* terms of like sign is $p^m(1-p)$. Thus the probability that a chosen term belongs to a run of exactly *k* terms is $kp^{k-1}(1-p)^2$. Hence in a large sample, if N is the total number of terms and R the total number of runs

$$R = \sum_{k=1}^{\infty} N p^{k-1}(1-p)^2$$

$$= N(1-p)$$

$$\therefore p = 1 - \frac{R}{N}$$

The investigator intending to use this method should therefore:-

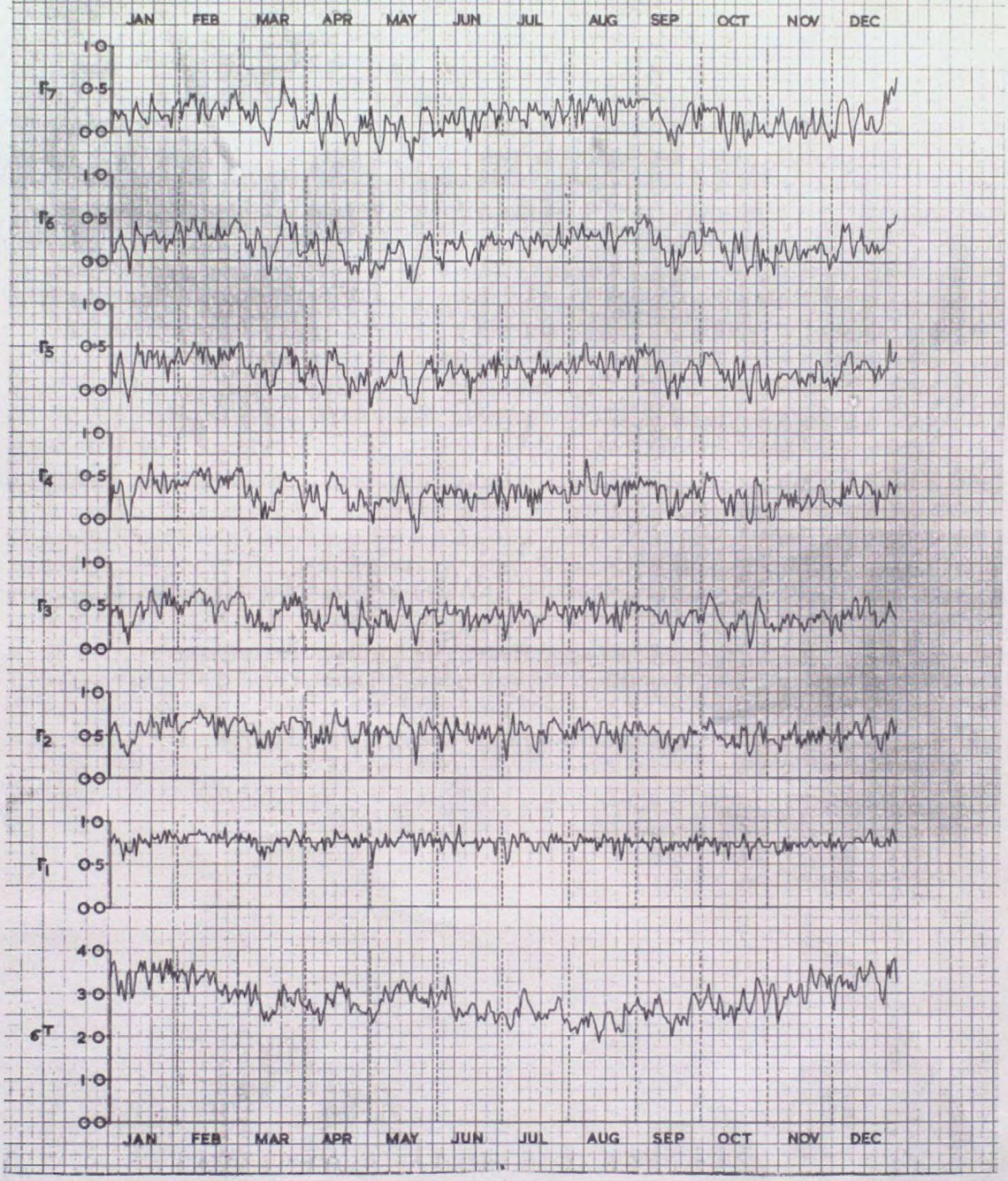
(i) satisfy himself that the distribution of the series studied does not depart seriously from normal

(ii) find the mean and then R and hence *p* by the above equation and

(iii) look up the corresponding value of *r* in Fig. 2.

"M.R.P. 1023"

FIG I. STANDARD DEVIATION OF THE MEAN DAILY TEMPERATURE AT KEW FOR EACH DAY OF THE YEAR (IN DEGREES CELSIUS) AND SERIAL CORRELATIONS BETWEEN THE TEMPERATURE ON EACH CALENDAR DAY AND THOSE OF THE SEVEN SUCCEEDING DAYS.



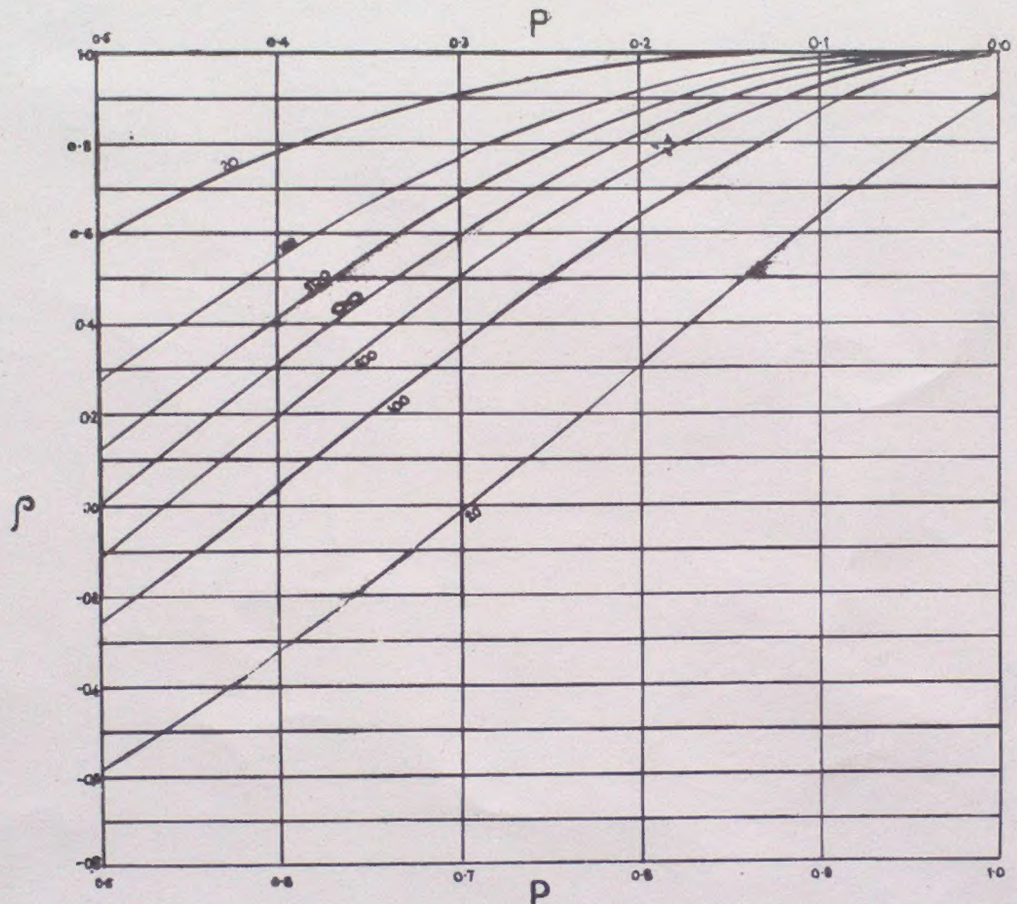


DIAGRAM FOR ESTIMATION OF CORRELATION COEFFICIENT WITH CONFIDENCE LIMITS